Approaches to Gene Mapping in Complex Human Diseases

Edited by

Jonathan L. Haines
Department of Molecular Physiology and Biophysics
Program in Human Genetics
Vanderbilt University School of Medicine
Nashville, TN

Margaret A. Pericak-Vance
Section of Medical Genetics
Center for Human Genetics
Duke University Medical Center
Durham, NC

WILEY-LISS

A John Wiley & Sons, Inc., Publication

New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ♾

Published simultaneously in Canada.

Library of Congress Cataloging in Publication Data:

Approaches to gene mapping in complex human diseases / edited by
Jonathan L. Haines and Margaret A. Pericak-Vance.
p. cm.
Includes bibliographical references and index.
ISBN 0-471-17195-6 (cloth : alk. paper)
1. Genetic disorders. 2. Human gene mapping. I. Haines,
Jonathan L., 1957– . II. Pericak-Vance, Margaret ann.
[DNLM: 1. Hereditary Diseases—genetics. 2. Chromosome Mapping.
3. Genetic Markers. 4. Genotype. QZ 50 A653 1998]
RB155.5.A67 1998
616'.042—dc21
DNLM/DLC
for Library of Congress 97-42666

Printed in the United States of America

10 9 8 7 6 5 4 3 2

Contents

15

Linkage Disequilibrium and Allelic Association

Margaret A. Pericak-Vance
Center for Human Genetics
Section of Medical Genetics
Duke University Medical Center
Durham, North Carolina

INTRODUCTION

There are two primary approaches for mapping genes that either cause or increase susceptibility to human disease. The first is the linkage analysis approach, either a parametric lod score analysis when the genetic model is known or model-independent affected relative pair analysis when the genetic model is unknown. The linkage approach has been discussed in detail elsewhere in this book (Chapters 12, 13, 14). The second approach is the application of allelic association studies. Allelic association is another nonparametric approach to mapping disease genes. It is a useful and often necessary tool in identifying disease gene loci, particularly susceptibility genes in genetically complex diseases. Allelic association can be explained either by direct biological action of the polymorphism (e.g., the *APOE-4* allele in Alzheimer disease), or by linkage disequilibrium with a nearby susceptibility gene. Association studies can play a critical role in the analysis of genetically complex traits, in the evaluation of candidate gene loci as well as in the fine-mapping of a region once linkage studies have indicated a region of interest for follow-up analysis. Certainly, as the Human Genome Initiative identifies and characterizes each gene in

Approaches to Gene Mapping in Complex Human Diseases, Edited by Jonathan L. Haines and Margaret A. Pericak-Vance. ISBN 0-471-17195-6

the human genome, this approach will become more widespread, particularly for localization and gene identification.

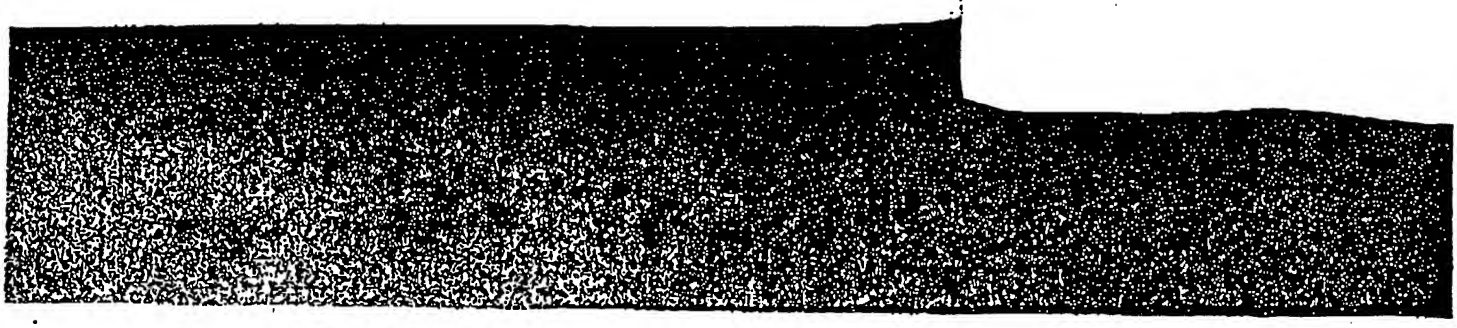
Allelic association refers to a significantly increased or decreased frequency of a marker allele with a disease trait and represents deviations from the random occurrence of the alleles with respect to disease phenotype. Allelic association can be due to either linkage or association. We use linkage disequilibrium to mean allelic association maintained by tight linkage. Linkage disequilibrium occurs when a particular marker allele lies so close to the disease susceptibility allele that these alleles will be inherited together over many generations. Thus the same allele will be detected in affected individuals in multiple apparently unrelated families. Conceptually this is the same as standard linkage analysis, except that the recombination distances being measured are now very small (generally < 1 cM), and the recombination events can only be inferred based on the level of sharing of the same allele. Population substructure most often occurs with the recent admixture of populations. In the case of population substructure, alleles may show a statistical association simply by chance due to differences in allele frequencies in the two mixing populations. This can occur even when there is no biological association or true genetic linkage.

LINKAGE DISEQUILIBRIUM

To understand one way in which linkage disequilibrium can come about, consider this hypothetical example. A new mutation occurs in a gene that results in a disease-causing phenotype. At the time of the initial mutation, every marker allele for every marker on the chromosome is "associated" with the disease mutation. The chromosome with the disease mutation is then transmitted to the offspring of the original individual in whom the mutation occurred. Transmission over several generations gives the opportunity for recombination to occur and thus for the rearrangement of the alleles at the marker loci. Alleles at marker loci that are further away from the disease mutation will exchange faster than markers that are closer to the disease mutation. The closer to the marker is to the disease gene, the longer the marker allele/disease association will persist.

When marker and disease loci are very close together on a chromosome, genetic crossing over will have occurred at such a low rate that the marker will appear to cosegregate with the gene regardless of the family studied. This is in contrast to the situation of two loci further apart but still linked, in which case repeated crossing over will allow all possible combinations of chromosomal haplotypes to appear with frequencies as predicted by the equation for Hardy–Weinberg equilibrium (Chapter 2). Thus, linkage disequilibrium can be very useful in defining the ancestral haplotype of a disease gene in relation to several marker loci; it can be used for fine-mapping of the disease gene even when complete linkage ($\theta = 0.0$) is established in the families being studied.

There are many measurements of linkage disequilibrium (Devlin and Risch, 1995). The most commonly used is the disequilibrium coefficient D.



$$D = P_{11} - p_1 q_1$$

where P_{11} is the observed frequency of the 1/1 haplotype, p_1 is the frequency of the "1" allele at locus 1 in the general population and q_1 is the population frequency of the "1" allele at locus 2. Generally, the "1" allele at each locus is defined as the most common of the alleles at that locus. Because we assign "1" as the most common allele, the coefficient D ranges from –0.25 to 0.25. Positive values of D indicate that the common alleles at each locus segregate together. Negative values indicate that the common allele at one locus segregates with the rare allele at the other locus. The rate of decay of linkage disequilibrium is dependent on the distance between loci:

$$D_t = D_0(1 - \theta)^t$$

where t is the current generation number, D_t is the current amount of disequilibrium, D_0 is the disequilibrium at generation 0, and θ is the recombination fraction between loci.

Allelic association due to population admixture, selection, or genetic drift between unlinked loci will decay fairly rapidly in comparison to linkage disequilibrium between tightly linked genetic loci, and thus is a short-term phenomenon that will be almost impossible to detect in the typical study. However linkage disequilibrium will decay rather slowly, dependent primarily on the recombination distance between the markers and the number of generations that has passed since the initial event (Fig. 15.1). The slowness of linkage disequilibrium decay makes this a useful mapping tool.

The general rule of thumb is that the stronger the disequilibrium, the closer the marker is to the disease locus. This is not always the case, however, for several reasons. First, the frequencies of the marker alleles have an impact on the power to detect linkage disequilibrium. For example, if the disease susceptibility allele is associated with a marker allele whose general population allele frequency is 0.50, an

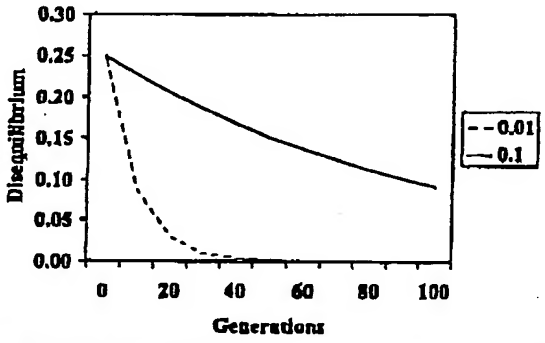


Figure 15.1 Decay in linkage disequilibrium for recombination fractions of 0.01 and 0.10.

observation that this marker allele frequency is 0.80 in affected individuals represents only a 60% increase in the frequency. However, if the marker allele population frequency is only 0.20, an observed 0.80 frequency is a 400% increase. Mutation rates at the marker locus also affect disequilibrium by increasing the chance that the associated marker allele will change and so seem to be representing a different chromosome.

Population bottlenecks, where the effective population size is reduced to a very small number for a period of time before the population size increases again, can create or reinforce an existing association. This is done by the random loss to the genome pool of most chromosomes carrying the susceptibility allele; what remains may have existed in only one individual who survived the bottleneck. Chance loss of susceptibility-allele-bearing chromosomes (random genetic drift) can also generate linkage disequilibrium. Two phenomena that can complicate the analysis of allelic association are selection in favor of a particular phenotype and new mutations (at either the disease or marker loci) arising in the population.

MAPPING GENES USING LINKAGE DISEQUILIBRIUM AND SPECIAL POPULATIONS

In most cases, allelic association will result from linkage disequilibrium, unless a specific polymorphism in the actual susceptibility gene is being studied. The power of linkage disequilibrium is best exploited in its use in fine-mapping. Because linkage disequilibrium rarely extends more than 1 cM from the susceptibility locus, its detection signals a significant decrease in the minimum candidate region. However, this great strength is also its great drawback. Because the effect is so localized, it will be very hard to find against the background of the entire human genome. In genetically complex diseases, the further complications of genetic heterogeneity and/or gene–gene interaction may make detection even more difficult.

Mapping and/or gene identification using linkage disequilibrium is especially powerful in genetically unique or isolated populations (so-called special populations, such as Amish or Finnish populations). These populations have already been used successfully for Mendelian diseases, since they are often homogeneous in disease origin. In other words, there are likely to be only a few founding individuals who carried specific chromosomal haplotypes on which the original mutation occurred. If the population is relatively new (i.e., the result of recent admixture of two populations), this approach can also be useful in the general mapping of disease loci as well, as has been shown in rare recessive disorders (Hastbacka et al., 1992). This is possible because the linkage disequilibrium is likely to extend over larger areas (several centimorgans) of the chromosome, since the number of generations available to allow decay of the linkage disequilibrium by recombination is small.

These populations can be equally useful in mapping complex traits. The basic premise is that genetically isolated populations will have fewer genes contributing toward a disease trait, and therefore the effect of each remaining susceptibility gene

will be easier to detect. The value of these populations in genetic mapping studies has long been realized. Homozygosity mapping was described early on in the molecular revolution (Lander and Botstein, 1987). Recently these advances have been expanded to include the use of pooling strategies (Sheffield et al., 1994) and the exploitation of the phenomena of linkage disequilibrium together with the isolated inbred nature of these groups through the approach of "shared segment" mapping of a complex phenotype (Houwen et al., 1994; Durham and Feingold, 1997). Thus the great advantage of the special population is its power to detect linkage. However, it must be pointed out that this power comes at the potential cost of specificity. Only one or a few of the entire suite of susceptibility genes may be found, and the effect of this gene or genes may be limited to the special population being studied.

ASSOCIATION STUDIES: IMPLEMENTATION

There are two types of association studies. *Case–control* studies compare allele frequencies in a set of unrelated affected individuals to a set of matched controls. The control populations should be matched with respect to ethnicity as well as other factors such as age. Spurious associations can result because of population stratification (i.e., the existence of multiple population subtypes in what is assumed to be a relatively homogeneous population). Such stratification can represent either recent admixture or the incorrect matching of cases and controls. The existence of these confounding factors can lead to a significant result even in unlinked loci (Table 15.1) or unassociated loci within stratum.

An example of a case–control linkage study is the identification of the *APOE-4* allele as the susceptibility gene in late-onset familial and sporadic AD. Table 15.2 presents data on a large study of over 500 Alzheimer disease patients and age and ethnically matched controls. Using standard chi-square analysis, a significant association was found [$p < 0.001$]. As indicated in Table 15.2, there appears to be an in-

Table 15.1 Example of Population Stratification[a]

	Population A			Population B			Population C (mixed)	
	A (0.80)	*a* (0.20)		*A* (0.20)	*a* (0.80)		*A* (0.50)	*a* (0.50)
B (0.80)	0.64	0.16	*B* (0.20)	0.04	0.16	*B* (0.50)	0.25	0.25
b (0.20)	0.16	0.04	*b* (0.80)	0.16	0.64	*b* (0.50)	0.25	0.25

[a]In this example populations A and B have very different allele frequencies for the disease gene *A* and the unlinked marker *B*. If the populations mix evenly, then the overall allele frequencies are as seen in population C. If comparisons of the haplotype frequencies are made from population C to either population A or B, the results will be significant even if no linkage exists. For example, if the affected individuals are drawn from population C, they will have an *A/B* haplotype frequency of 0.25, assuming no association with the disease. If the controls are drawn from population B the *A/B* frequency will be 0.04. The erroneous conclusion is that there is an association of the disease with the *B* allele (0.25 vs. 0.04), since they occur together more often than in population B. In most data sets, it is not possible to determine whether the sample was drawn from one, two or more populations.

Table 15.2 Case-Control Association Studies: APOE-4 allele and Alzheimer Disease

a. Observed Counts

APOE-4 allele	Cases	Controls	Total
APOE-4	240	60	300
Not APOE-4	360	340	700
Total	600	400	1000

b. Expected Counts

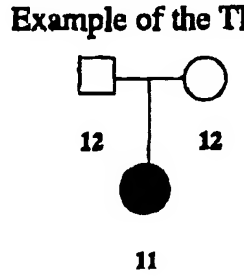
APOE-4 allele	Cases	Controls	Total
APOE-4	180	120	300
Not APOE-4	420	280	700
Total	600	400	1000

χ^2 = (observed-expected)^2/expected = ((240 − 180)^2/180) + ((60 − 120)^2/120) + ((360 − 420)^2/420) + (340 − 280)^2/280)) = 71.5. $P < 0.0001$.

crease in the *APOE-4* allele, and a concomitant decrease in the *APOE-3* allele, in Alzheimer patients.

Family-based studies control for the possibility of genetic differences between the case and control populations by comparing the frequencies of alleles transmitted to the affected child to the alleles not transmitted. The only samples necessary are those from the affected individual and his or her two parents (the TDT triad). This approach eliminates the concern that population substructure may be the cause of the association. These studies include the transmission disequilibrium test (TDT) (Spielman et al., 1993), the haplotype relative risk test (HRR) (Falk and Rubenstein, 1987), and the AFBAC method (Thomson, 1995). The HRR and AFBAC approaches were developed as family-based tests for association, and the AFBAC method is designed to detect association in the presence of linkage. The TDT approach tests for linkage in the presence of association. Both AFBAC and TDT have little power unless linkage and association coexist. The difference between these two methods is that the TDT can also function as a test of association in the presence of population admixture and can be used as a valid test of linkage. The statistical differences between these methods are subtle and are not described here.

The TDT is the most widely used of all the tests. It can be a more powerful test to detect linkage than the affected sib pair linkage tests, especially when the genetic effect is small (Spielman et al., 1993; Risch and Merikangas, 1996), as is often the case with genetically complex traits. The disadvantage of this approach is that the TDT has no power to detect linkage if association is not present. The TDT test examines the number of transmissions of allele 1 (*A1*) or allele 2 (*A2*) from a heterozygous parent to an affected offspring. An example of the TDT is given in Figure 15.2. As a test of linkage the counts needed in Figure 15.2 can come from simplex, multiplex or even multigenerational family data. The statistical significance is tested by standard chi-square (McNamar's test). The data used in con-



Transmitted	Not Transmitted	
	Allele 1	Allele 2
Allele 1	0	2
Allele 2	0	0

Figure 16.2 Example of scoring a transmission disequilibrium test (TDT) family.

structing the counts comes only from heterozygous parents. The TDT statistic is as follows:

$$\chi_1^2 = \frac{(b-c)^2}{b+c}$$

where b is the number of times an *A1/A2* parent transmits an *A1* to an affected offspring and c is the number of times an *A1/A2* parent transmits an *A2* to an affected offspring.

The test statistic then tests for deviations from the expected equal transmission rate into the two categories from the heterozygous parents. Homozygous parents do not need to be scored. This is different from the AFBAC method, which uses both heterozygous and homozygous parents. A significant result indicates that the marker is linked to the disease locus. The TDT can find linkage only in the presence of association. If there were only linkage and no linkage disequilibrium, then across families there would be no difference between b and c, since the allele in coupling with the disease gene in each family is random, preventing the detection of linkage. The TDT can also serve as a test of linkage disequilibrium if only simplex families are used or if only one affected individual and his or her parents are included per family. This use of the TDT is critical for narrowing a broad region of interest identified by linkage analysis. Analysis of association could potentially identify the markers that are closest to the actual disease susceptibility locus.

The TDT approach was a useful approach in identifying the relationship of the

Table 15.3 Transmission Disequilibrium Test (TDT) and Diabetes[a]

	Nontransmitted[b]	
Transmitted	*A1*	*A2*
A1	NI	78 (62)
A2	46 (62)	NI

[a]In this example, the number of *A1* alleles transmitted from a heterozygous parent to the affected (IDDM) child is 78, while the number of *A2* alleles transmitted is 46. If no association exists, the expectation is that each allele would be transmitted 62 times. This result is highly significant.
[b]NI, not informative; expected number in parentheses.
Source: Spielman et al. (1993).

insulin gene in insulin-dependent diabetes mellitus (IDDM) (Spielman et al., 1993) (Table 15.3). The TDT method does not require researchers to go to the expense and effort of recruiting families with multiple affected individuals, indeed, the clinical status of the parents does not have to be known. In some cases, however, these advantages may be outweighed by practical problems. In the diabetes situation, both parents of the affected family members were available for study, making this approach ideally suited to IDDM. However, in the Alzheimer/*APOE-4* example, since AD is a late-onset disease, parental DNA is almost always unavailable, and thus the traditional TDT approach was impossible. A novel approach that circumvents this difficulty is to use unaffected siblings as controls rather than relying on parental controls (Curtis, 1997; Boehnke and Langefeld, 1998; Spielman and Ewens, 1998). This sib-TDT (S-TDT) approach compares marker allele frequencies in affected and unaffected siblings. The test requires only a simple affected/unaffected sibling pair, although power can be increased somewhat if additional siblings are available. The S-TDT retains the advantages of the TDT in that it provides a test for linkage and association and is immune to the effects of sampling bias.

The TDT test was originally developed to look at biallelic marker systems or situations of alleles that could be readily collapsed because of prior information regarding a known association. With the availability of a multitude of multiallelic markers, several new statistics have been proposed. These included the symmetry statistic (T_c), the marginal statistic (T_m) of Bickeboller and Clerget-Darpoux (1995), the likelihood ratio statistic (T_l) of Sham and Curtis (1995), and the marginal statistic with only heterozygous parents (T_{mhet}) of Spielman and Ewens (1996). Kaplan et al. (1997), investigated the properties of these four statistics and determined that the T_{mhet} was the most appropriate and efficient method. It had equivalent power to the other tests and it gave a (approximately) valid chi-square test of linkage. Their recommendation for multiallelic markers was to implement the T_{mhet} using critical values of χ^2 with $(m - 1; m =$ number of alleles at the marker locus) df and including all available affected individuals and their heterozygote parents in the analysis. Table 15.4 provides the definition of the T_{mhet} statistic.

Table 15.4 Multiallelic TDT: T_{mtmt}*

Nontransmitted allele	Transmitted allele			
	1	2	3	Total
1	n_{11}	n_{12}	n_{13}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	$n_{3.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{..}$

*$T_{mtmt} = \frac{m-1}{m} \sum_{i=1}^{m} \frac{(n_{i.} - n_{.i})^2}{n_{i.} + n_{.i} - 2n_{ii}}$ where m is the number of alleles.

THE USE OF THE TDT IN GENOMIC SCREENS

The TDT was originally proposed as a test for linkage to specific candidate loci. More recently discussions have centered on the use of the TDT approach for entire genome scans (Risch and Merikangas, 1996). The ability to replicate in a comparable data set is as critically important when one is using the TDT as it is in the performance of linkage studies using affected relative pair analysis. With the advent of microchip technology on the horizon, and the interest in the development of ancestral single nucleotide repeat polymorphisms (SNPs), the expanded use of this methodology to quickly map and identify genes is a certainty.

However, there are several potential problems with the TDT in genomic scanning. First, the problem of multiple comparisons arises in this situation. That is, when so many statistical tests are performed false positive results are likely by chance alone unless the usual significance value of 0.05 or 0.01 is modified. Thus use of a critical value that is greater than the nominal P value is warranted. The usual Bonferroni correction approach (simply dividing the desired nominal significance level by the number of tests performed) will be too conservative because it assumes each test to be independent of the others. This will not be the case, since many of the markers will be linked and associated with each other. Unfortunately, it is not clear what the appropriate correction needs to be, although simulation-based statistics are being explored.

The second problem is simply the number of polymorphic markers necessary. Even at one marker per centimorgan, over 3000 well-mapped markers are needed. In addition, these markers must actually be located at 1 cM distances, not just scattered with a 1 cM average distance. While many more than 3000 markers exist, there are still many regions of the genome up to 20 cM in length with no known polymorphisms.

The third problem is that the TDT approach rests completely on the assumption that some level of linkage disequilibrium exists. While this may be true in many cases, susceptibility alleles arising from frequent mutation events, existing as ex-

tremely old mutations, or arising in regions with very high recombination rates, will have little or any detectable linkage disequilibrium.

SUMMARY

Allelic association may arise for several reasons, but association due to linkage disequilibrium can be exploited to aid in the mapping of genetically complex diseases. Both case–control and family-based methods can be used. The latter have several advantages, especially when tested using the TDT or its variant, the sib-TDT. The TDT is both simple and powerful, and it will have substantial power for detection of susceptibility alleles as better and more finely spaced markers are available for study.

REFERENCES

Bickeboller H, Clerget-Darpoux F (1995): Statistical properties of the allelic and genotypic transmission/disequilibrium test for multiallelic markers. *Genet Epidemiol* 12:865–870.

Boehnke M, Langefeld CD (1998): Genetic association mapping based on discordant sib-pairs: The discordant alleles test (DAT). Submitted.

Curtis D (1997): Use of siblings as controls in case control association studies. *Ann Hum Genet* 61:319–333.

Devlin B, Risch N (1995): A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 29:311–322.

Durham LK and Feingold D (1997). Genome scanning for segments shared identical by descent among distant relatives in isolated populations. *Am J Hum Genet* 61:830–842.

Falk CT, Rubinstein P (1987): Haplotype relative risks: An easy way to construct a control sample for risk calculations. *Ann Hum Genet* 51:227–233.

Hastbacka J, de la Chappelle A, Kaitila I, Sistonen P, Weaver A, Lander E (1992): Linkage disequilibrium mapping in isolated founder populations: Diastrophic dysplasia in Finland. *Nat Genet* 2:204–211.

Houwen R, Baharloo S, Blankenship K, Raeymackers P, Juyn J, Sandkuijl LA, Freimer NB (1994): Genome screening by searching for shared segments: Mapping a gene for benign recurrent intrahepatic cholestasis. *Nat Genet* 8:380–386.

Kaplan NL, Martin ER, Weir BS (1997): Power studies for transmission/disequilibrium tests with multiple alleles. *Am J Hum Genet* 60:691–702.

Lander ES, Botstein D (1987): Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. *Science* 236:1567–1570.

Risch N, Merikangas K (1996): The future of genetic studies of complex human disorders. *Science* 273:1516–1617.

Sham PC, Curtis D (1995): An extended transmission/disequilibrium test (TDT) for multiallele marker loci. *Ann Hum Genet* 59:323–336.

Sheffield V, Carmi R, Kwitek-Black A, Rokhlina T, Nishimura D, Duyk GM, Elbedour K, Sunden SL, Stone EM (1994): Identification of a Bardet-Biedl syndrome locus on chro-

mosome 3 and evaluation of an efficient approach to homozygosity mapping. *Hum Mol Genet* 3:1331–1335.

Spielman RS, Ewens, WJ (1996): The TDT and other family-based tests for linkage disequilibrium and association. *Am J Hum Genet* 59:983–989.

Spielman RS, Ewens WJ (1998): A sibship test for linkage in the presence of association: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* (in press).

Spielman RS, McGinnis RE, Ewens WJ (1993): Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–516.

Terwilliger JD, Ott J (1992): A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* 42:337–346.

Thomson G (1995): Analysis of complex human genetic traits: An ordered-notation method and new tests for mode of inheritance. *Am J Hum Genet* 57:474–486.